MoralStrength: Exploiting a moral lexicon and embedding similarity for moral foundations prediction☆

Oscar Araque [a,*,1], Lorenzo Gatti [b,2], Kyriaki Kalimeri [c,3]

[a] *Intelligent Systems Group, Universidad Politécnica de Madrid, Madrid, Spain*
[b] *Human Media Interaction Lab, University of Twente, Enschede, The Netherlands*
[c] *Data Science & Digital Philanthropies Laboratory, ISI Foundation, Turin, Italy*

ABSTRACT

Moral rhetoric plays a fundamental role in how we perceive and interpret the information we receive, greatly influencing our decision-making process. Especially when it comes to controversial social and political issues, our opinions and attitudes are hardly ever based on evidence alone. The Moral Foundations Dictionary (MFD) was developed to operationalize moral values in the text. In this study, we present *MoralStrength*, a lexicon of approximately 1,000 lemmas, obtained as an extension of the Moral Foundations Dictionary, based on WordNet synsets. Moreover, for each lemma it provides with a crowdsourced numeric assessment of *Moral Valence*, indicating the strength with which a lemma is expressing the specific value. We evaluated the predictive potentials of this moral lexicon, defining three utilization approaches of increased complexity, ranging from lemmas' statistical properties to a deep learning approach of word embeddings based on semantic similarity. Logistic regression models trained on the features extracted from *MoralStrength*, significantly outperformed the current state-of-the-art, reaching an F1-score of 87.6% over the previous 62.4% (*p*-value < 0.01), and an average F1-Score of 86.25% over six different datasets. Such findings pave the way for further research, allowing for an in-depth understanding of moral narratives in text for a wide range of social issues.

1. Introduction

Language usage reflects our thoughts, emotions, values, and culture, as we communicate with others. With the burst of on-line communication and social media, people are empowered to express and broadcast their opinions on contentious issues, timely, and at greater scale. This unprecedented opportunity allows scientists and policymakers to study phenomena such as opinion formation, radicalization, and polarization in society, as they happen.

In this study, we propose a lexicon for detecting and quantifying the moral rhetoric behind people's judgments, as reflected in spontaneous digital interactions. Moral values influence the way we rationalize and take a stance upon controversial topics, like abortion [1], homosexuality [1], climate change [2], poverty [3] or health related attitudes like for instance exercising [4] and vaccine advocacy [5,6]. They are also closely related to our political views [7] and the opinion formation mechanisms regarding immigration [8], political extremism [1,9,10], and intimate partner violence [11]. Recently, scientists also showed that moral values could be employed to detect violent protests [12] based on user-generated text.

We operationalize morality via the Moral Foundations Theory (MFT) [13], which expresses the psychological basis of morality in terms of innate intuitions, defining the following five foundations: *care/harm*, *fairness/cheating*, *loyalty/betrayal*, *authority/subversion*, and *purity/degradation* (see [14,15]). Even if in its infancy, MFT is the most well-established theory in the psychology and social sciences. It is also broadly adopted in the computational social science field since it defines a clear taxonomy of values together with a term dictionary, the Moral Foundations Dictionary (MFD, hereafter) [13], which is an essential resource for natural language processing applications. The creators of the MFD, highlight the difficulty of creating such a resource since linguistic, cultural, and historical context reflect on language usage. Among the most significant limitations of the MFD, we have: (i) a limited amount of lemmas and stem of

---

words; (ii) "radical" lemmas rarely used in everyday language, for instance, "homologous" and "apostasy"; and (iii) an association with a moral bipolar scale, so-called vice and virtue, but without any indication of "strength".

Here, we address precisely these shortcomings; initially, we expanded the existing MFD using the WordNet lexical database [16] and then, we provide a set of normative ratings for empirical assessment of morality. The resulting lexicon, namely *MoralStrength*, offers approximately three times more lemmas while going beyond the binary nature of the MFD. Moreover, we present a machine learning framework exploring the potentials of *MoralStrength* in predicting the moral narratives from the user-generated text.

The suggested framework includes three models of increasing complexity; two of them are based on straightforward feature extraction from lemma frequencies and statistical properties, while the third one is based on embedding representation of semantic similarity. We thoroughly evaluate the proposed framework employing the Moral Foundations Twitter Corpus (MFTC) [17]. The MFTC corpus is a collection of seven Twitter dataset previously employed in studies related to moral detection from text. It consists of approximately 35,000 tweets along with their respective annotations according to the MFT foundations regarding critical social issues.

Importantly, the performance of our approach in predicting morality is outperforming the current state-of-the-art methods. Our results show that the pure textual representations emerged from the *MoralStrength* lexicon greatly benefit the performance of the prediction. These findings pave the way towards a more in-depth understanding of moral judgments, dispositions, and attitudes formation from spontaneous digital data.

Hence, we contribute to the research and policymaking communities with a useful resource and a concrete framework that can be employed for analyzing large scale user-generated communications, or even nowcasting people's attitudes and opinions on controversial phenomena. When it comes to critical social issues, the proposed approach can provide insights to understand personal narratives and viewpoints better, but also how people will potentially reason on the information they receive. Such knowledge is essential for policymaking specialists to design effective communication campaigns that appeal to people's values, given the ever-increasing penetration of social media to the population.

2. Related literature

Psychologists and social scientists have systematically analyzed text data to address their research questions. Back in the late '60s, the Harvard General Inquirer dictionary [18] was the precursor of sentiment analysis which was to become, together with opinion extraction, a core theme of natural language processing (NLP). Ever since scientists gradually increased complexity moving from simple techniques (e.g. the unsupervised and partially rule-based approach of [19]) to sophisticated methods that try to determine the context of words (e.g. [20]). The topics addressed also became more challenging tackling notions such as irony and sarcasm [21]. With time, not only the methods became more sophisticated but also the tasks become more ambitious. Extensive studies on linguistic markers of sentiment and affect [20,22–25] paved the way to assess more complex constructs such as personality [26,27] and human values [28,29]. Moral values are considered to be a higher level construct with respect to personality traits, determining how and when dispositions and attitudes relate with our life stories and narratives [30]. Here we provide a brief literature review of the studies regarding moral values assessment from textual data. As in sentiment and

personality analysis, also, in this case, pioneer works followed a dictionary-based approach, while the current state of the art performance is based on deep learning.

The first vocabulary developed to assess the moral values from textual data was the Moral Foundations Dictionary (MFD) [13]. It was used together with the Linguistic Inquiry and Word Count (LIWC) software [31] to estimate moral traits and to investigate differences in moral concerns between different cultural groups. Clifford et al. [32] employed the MFD for performing manual text analysis of 12 years of coverage in the New York Times focusing on political debate in the US. Teernstra et al. [33] assessed the political debate regarding the "Grexit" from approximately 8000 tweets. They compared the performance of using the raw data, bigrams, and the MFD features in employing basic machine learning models, namely, Naive Bayes (NB) and Maximum Entropy (ME). They concluded that pure machine learning is preferable to dictionary approaches since it has similar prediction accuracy while using fewer assumptions. In this study, we follow a similar approach to [33]; however, we propose an expanded version of the MFD, including also the moral valence per lemma. Moreover, we employ logistic regression models to infer moral values from uni-grams combined with lexical features.

Dehghani et al. [34] examined the differences between liberal and conservative moral value systems using a hierarchical generative topic modeling technique, based on Latent Dirichlet Allocation (LDA) [35], to enable the unsupervised detection of topics in their corpus of liberal and conservative weblogs. They used small sets of words selected from the MFD as seeds to encourage the emergence of issues related to different moral concerns and examined similarities and differences in how such matters are expressed between these groups. Consistently with findings in moral psychology, they demonstrate that there are significant differences in how liberals and conservatives construct their moral belief systems. Sagi et al. [36] employed the same framework to study moral rhetoric in text for a specific case study, the US Federal shutdown of 2013 [1], where they examined the role of morals in intra- and inter-community differences of political party retweets. In both works, they were based on the framework presented in [34], where LDA was employed to create a co-occurrence matrix on which the similarity between the texts and the vectors representing the different MFT moral traits was computed. In a similar approach, Kaur et al. [37] attempted to quantify the moral loadings of text, based on the Latent Semantic Analysis (LSA). They used a bag-of-words model, representing the entire corpus by a word-context matrix. Then they reduced its dimensionality obtaining low-dimensional word vectors, in which similar vectors represent similar meaning words. Our study is presenting a different approach since we do not use LSA representations, but rather pre-trained word embeddings models. Although pre-trained word embeddings do not contain domain-specific knowledge, they express language regularities encoded as offsets in the resulting vector space. The proposed representations based on the work of Araque et al. [38], exploit precisely the similarity between the analyzed text and a selection of words with moral content.

More recently, Garten et al. [39] employed the MFD to detect moral rhetoric in general, and more specifically, shifts in long political speeches over time. Then, based on psychological dictionaries and semantic similarity to quantify the presence of moral sentiment around a given topic, Garten et al. [40], proposed the Distributed Dictionary Representations (DDR) method. Showing promising results, DDR was also employed by Hoover et al. [41] to detect moral values in charitable giving. Later on, Garten et al. [42] extended the method, incorporating demographic embeddings into the language representations. Our approach is based on an expanded version of the MFD, with evaluated manual annotations regarding the moral valence of each lemma, that can be incorporated in computational frameworks.

In a study more similar to ours, attempting to predict moral values involved in Twitter posts, Lin et al. [43] proposed a method that automatically acquires background knowledge to improve the moral value prediction, pointing out the difficulty of the task also for human experts. Based on the work of [12,13,43] predicted the moral sentiment of the tweets. Their model consists of three layers, an embedding (lookup) layer, a recurrent neural network (RNN) with long short-term memory (LSTM) [44] and an output layer. The first layer converts words in an input tweet to a sequence of pre-trained word embeddings, the LSTM layer processes these embeddings and outputs a fixed-sized vector which encodes critical information for moral value prediction, while a vector representing the percentage of words that match each category in the Moral Foundations Dictionary [13] are concatenated with the LSTM feature vector. Our approach is differentiating to this one since we employ word embeddings to compute the similarity between words rather than directly feeding them in a neural network architecture. Along the same line, Rezapour et al. [45] investigated the relationship between moral values and stance towards a series of social issues. Their findings show that enhancing the original MFD improves the prediction accuracy of morality in text. They expanded the original MFD and employed a series of machine learning classifiers (SVM, RF and LSTM) predicting the moral traits. This study underlines the importance of expanding the MFD; we go one step further introducing the notion of moral "strength" while showing how more abundant information is improving the overall accuracy of the models.

The core contribution of this study is the extended lexicon of moral lemmas with the respective moral valence. To explore the properties and full potentials of the lexicon, we suggest three different models of increasing complexity, demonstrating the value of this resource. The proposed approaches range from feature engineering methods to a system which employs word embeddings of semantic similarity based on the work of Araque et al. [38].

3. Materials and methods

3.1. Expansion of the moral foundations dictionary (MFD)

The cornerstone of our study is the Moral Foundations Dictionary (MFD) [13] which was created to capture the moral rhetoric according to the five predefined dimensions defined by the Moral Foundations Theory (MFT) [46]. The original MFD[4] consists of lemmas and stems divided into "virtue" and "vice" [13] for each foundation according to their moral polarity. "Virtue" words are foundation-supporting words (e.g., *safe** and *shield* for Care "virtue"), whereas "vice" words are foundation-violating words (e.g., *kill* and *ravage* for Care "vice"). MFD [13] was meant to be used together with the Linguistic Inquiry and Word Count (LIWC) program [47], and thus contains either lemmas (158 entries such as *abandon*) or stems with a wild-card sign (166 listings), that LIWC analysis uses to match with all the forms of the base word; for instance, the entry *abuse** will match "abuse", "abuses", "abused", "abuser", "abusers", and so on. Due to the limited amount of lemmas and stems of words, often radical or rarely used in everyday language, for instance, "homologous" and "apostasy", the expansion of the existing dictionary is of essential importance (see Fig. 1).

Since we are interested in lemmas instead of stems, we initially expanded the original dictionary using the WordNet [16] synsets, maintaining the lemmas that shared the same initial part with stems in the MFD. The result of this first expansion was to obtain for each MFD entry, for instance, *traitor**, a series

Table 1
Corpus size after employing the WordNet resource to expand the MFD according to the official "virtue" and "vice" categories. The initial number of words contained in the MFD is shown in parenthesis.

Moral Dimension	Virtue	Vice
Care/Harm	95 (16)	85 (35)
Fairness/Cheating	69 (26)	57 (18)
Loyalty/Betrayal	99 (29)	72 (23)
Authority/Subversion	160 (45)	101 (37)
Purity/Degradation	97 (35)	161 (55)
Total	520 (151)	476 (168)

of lemmas, for instance, *traitor#n*, *traitorous#a*, *traitorously#r*, *traitorousness#n*.[5]

We performed an initial preprocessing step on the obtained word corpus removing the forms that matched the search but did not relate to a moral trait. For example, the stem *caste** not only matches *caste#n* and *caste_systems#n*, but also *caster#n* and *caster_sugar#n*, which are clearly not related to any moral foundation. This procedure was carried out manually, considering both the gloss for the lemma provided by WordNet and the moral trait that should be attributed to that word (e.g., while it could be argued that a statesman' name is an appropriate match for the Authority trait, the stem *church** relates to purity, and thus we ignored *Churchill#n*).

Following the original classification, we divided the obtained word corpus (1148 words) in "virtue" and "vice" lemmas resulting with 520 "virtues" and 476 "vices" while 152 were characterized as "general" morality words. These words can pertain to more than one traits; however, this is not common as the dataset consists of 442 unique "vice" words and 512 unique "virtue" words as shown in Table 1.

3.2. Moral valence annotation

Once the expanded dictionary was obtained, we used the Figure Eight[6] crowdsourcing platform to annotate each lemma with an association strength to the related moral trait. The goal here is twofold. On the one hand, we can use these annotations to determine if the terms extracted during the expansion process are still related to a moral trait. On the other hand, a lexicon with ratings could be useful for better dictionary-based approaches and is a first step in the direction of moral detectors that can rank sentences, instead of merely classifying them with a binary vice/virtue rating.

The expanded dictionary was annotated in terms of moral valence, but we also collected ratings of valence and arousal, following the definitions employed for the ANEW resource [16]. For our purpose, moral valence can be represented by a bipolar scale that, in aggregate, defines a continuous dimension from one moral extremity of the MFT to the other, e.g., from Care to Harm. Moral valence was operationalized in a 9-point Likert Scale, wherein if a word was ranked in the middle of the scale, it was semantically neutral to the specific moral dimension. The annotators were presented with the description of the moral trait and were asked to rate the relevance of the word to the specific foundation; if relevant, they were asked to rate its emotional valence, arousal and then its moral valence. Each experiment presented 20 different words to the annotator. The first time the annotators participated in the rating of a specific moral dimension (e.g., Care/Harm), after the experiment, they were asked to

---

[5] The letter after the number sign # indicates the part of speech for that word, i.e., *#n* for nouns, *#a* for adjectives, *#r* for adverbs, and *#v* for verbs.

[6] The Figure Eight Platform is available here: https://www.figure-eight.com/.
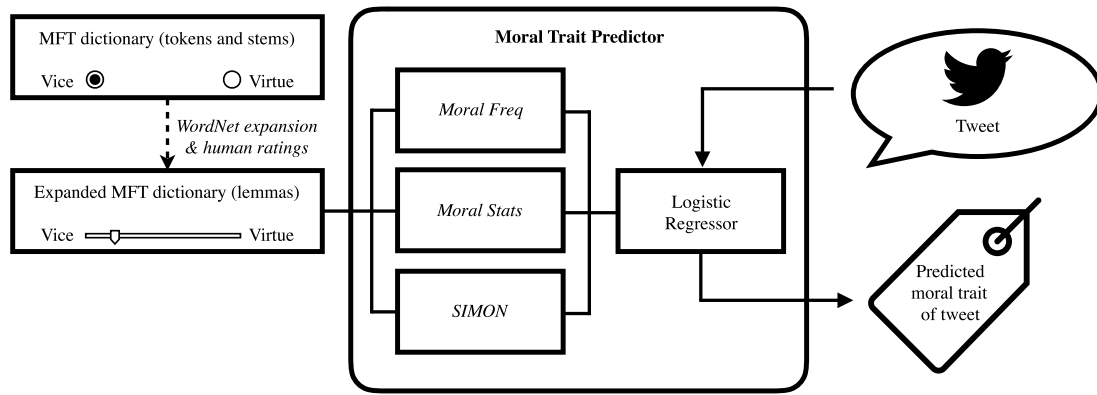
Fig. 1. Overview of the process, from dictionary expansion to moral value prediction.

fill in the Moral Foundations Questionnaire [13] for the respective dimension. At least five annotators were recruited for each lemma.

The ratings of valence and arousal were included to ensure a minimum quality of the annotation. Since no existing resource annotates moral valence on a fine scale, we used the values of valence from the subset of words that appear both in our extended dictionary and in [48]. Annotators have always been presented 4 "gold" words among the 20 words they annotate, and the annotations of those who fail more than 1 gold word are discarded. A valid answer is one that lies within 1.5 standard deviations from the valence mean of [48], for each specific gold word.

As seen, the proposed lexicon has both subjective and generative components. We need to take into consideration the subjectivity of human annotators; still, the candidate words for the annotations were chosen automatically, as a result of the expansion from a large word seed.

3.3. Moral lexicon approaches

For the generated moral lexicon, we propose the following feature extraction approaches, which can be divided into those that solely exploit the semantic information of each word, and those who exploit the moral valence associated to the word. More specifically, we propose three lexicon utilization approaches: (i) frequency counts, (ii) statistical summary, and (iii) word embedding similarity based representations. The two first approaches use both the words and their moral values, while the third one makes use solely of the selection of words, ignoring the associated numeric moral values. Fig. 2 illustrates these approaches. These three methods described above are used as feature extractors. In the next experiments, we feed these features to a logistic regression classifier. Such a simple learning algorithm is used to evaluate the performance of the proposed features, without exploiting more complex learning methods.

Moral Freq. It consists of counting the number of words that express a specific moral dimension in a binary way. To decide if a specific word expresses a moral, we apply a simple rule: if the word has its moral value lower than a certain threshold, it does not convey that moral; if higher, the word does express that moral. Given the properties of the generated moral lexicon, the threshold is set at 5. We represent a given text with a 10-dimensional vector, which contains the corresponding normalized frequency counts, each for each moral extremity; for instance, *care/harm* are represented by two dimensions, one for *care* and other for *harm*.

Moral Stats. Given a specific text, we generate a statistical summary of the moral valence distribution of the contained words in the text. In the statistical summary, we included (i) the average, (ii) the standard deviation, (iii) the median, and (iv) the maximum value. As a result, the text is represented by a 20-dimensional vector consisting of the statistical values obtained from the lexicon annotations.

SIMilarity-based sentiment projectiON (SIMON). Finally, the third approach is known as SIMilarity-based sentiment projectiON (SIMON), described in [38]. This method was initially developed for sentiment analysis tasks, while here, we adapted it to moral valence assessment. *SIMON* uses a pre-trained word embedding model to compute the cosine similarity between the words of the analyzed text and a selection of domain-related words, in our case, a specific moral dimension. Projecting the analyzed text over the selection of words from *MoralStrength*, we result with a vector representation that encodes the similarity of the document to the specific moral dimension.

3.4. Data collection and preprocessing

For the evaluation of our models, we employed the Moral Foundation Twitter Corpus (MFTC) [17]. MFTC is the most extensive available corpus containing 35,108 tweets and annotations, specifically collected to assess the moral values from user-generated content. This corpus includes seven distinct datasets[7] which were employed in scientific studies to assess the moral narratives in the user-generated text according to the moral foundations' theory. Here we provide an overview of the datasets that are included in the MFTC and are employed in our analysis.

Hurricane Sandy (HS). The first dataset we employed is presented in [43] and originally consisted of 4191 tweets.[8] These Tweets contain hashtags relevant to the "Hurricane Sandy", a hurricane that caused significant damage to the Eastern seaboard of the United States in 2012. Due to Twitter regulations, the original dataset could not be fully recovered, leaving us with only 3853 messages. We further removed the retweets, keeping only the original messages, to avoid overfitting the data. In this way, the processed dataset consists of 3478 instances.

Baltimore Protest (BP). The second dataset is comprised of messages related to the 2015 Baltimore Protests, which were motivated by the death of Freddie Gray. An older version of this dataset exists which contains a more significant number of instances. Nevertheless, since the annotations of this older version are obtained by automated means [12], we have decided to use the newer version, which has manual annotations.

---

7 The full dataset is available at https://osf.io/k5n7y/.
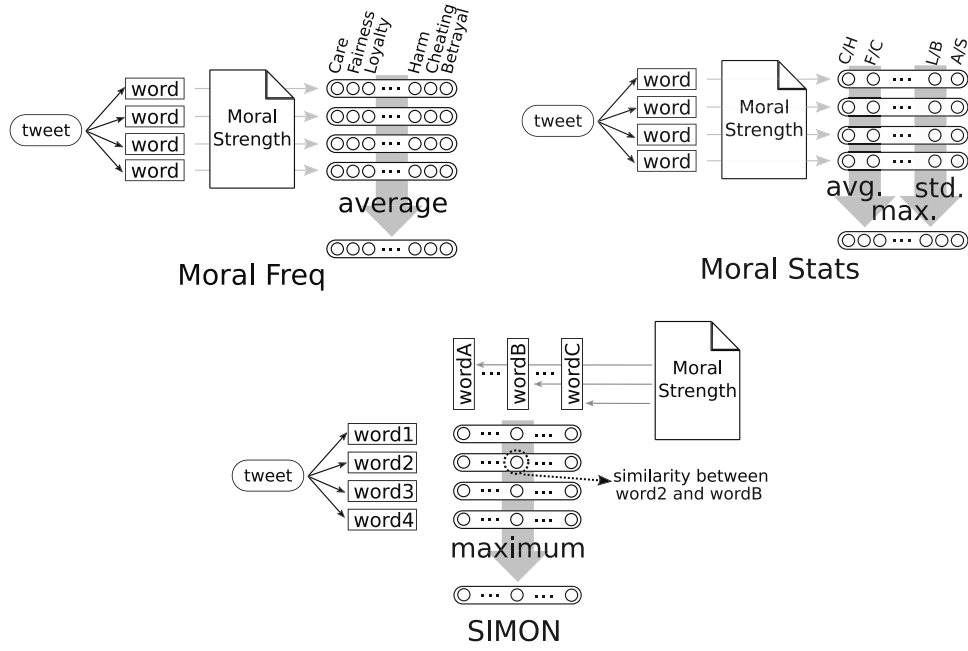8 This dataset can be obtained from https://osf.io/nzx3q/.

Fig. 2. Diagram of the proposed feature extraction approaches that utilize the presented lexicon.

All Lives Matter (ALM). Include #BlueLivesMatter and #AllLivesMatter hashtags and were posted between 2015–2016. These tweets were purchased from a third-party vendor.

Black Lives Matter (BLM). Posted between 2015–2016 about the Black Lives Matter Movement. Hashtags used to compile the corpus: #BLM, #BlackLivesMatter. The tweets were purchased from a third-party vendor.

2016 Presidential Election (PE). Scraped during the 2016 Presidential election season from the followers of @HillaryClinton, @realDonaldTrump, @NYTimes, @washingtonpost, and @WSJ.

Davidson (D). Taken from Davidson et al.'s [49] corpus of hate speech and offensive language.[9]

All the above datasets are annotated by experts who indicated the presence or absence of a moral foundation dimension for each tweet. Moreover, annotations include a "non-moral" label, indicating that the specific text does not reflect any moral trait. Importantly, the above datasets cover a wide variety of topics, both political and not. Topics related to politics cover left (e.g. BLM), right (e.g. ALM), and bilateral sides (e.g. Presidential Elections); while the datasets that are not unrelated to politics are expressing two controversial situations, a humanitarian call (e.g., Hurricane Sandy) and a collection of hate speech (e.g. Davidson). Such variability in the topics allows for a broader evaluation of the models' performance, avoiding biases due to context-specific language usage.

All data were collected downloading the original tweets following the Twitter IDs provided in the MFTC [17]. Since users often delete their tweets, we only managed to recover a portion of the original datasets. More specifically, 82% of the original dataset has been recovered, and the statistics are reported in Table 2. We also report the distribution of Tweets per moral dimension per dataset. We applied some basic preprocessing to the original textual content of the tweets employing the GSITK library.[10] In particular, we normalized the text converting the URLs using the special token "<url>", usernames to "<username>", and hashtags to the token <hashtag> and the word that is

Table 2
Statistics of the Moral Foundations Twitter Corpus employed in this study as a benchmark. All datasets were annotated by human annotators. Moreover, we note that according to the topic the distribution of the traits is varies. Last row reports the average number of training instances when using under-sampling.

Moral	Sandy	Baltimore	ALM	BLM	Elections	Davidson
Care	217	434	1314	1065	798	462
Fairness	416	292	723	940	736	133
Loyalty	410	895	408	531	286	331
Authority	155	120	274	494	177	1064
Purity	38	37	182	254	349	122
No moral	2242	2396	585	1056	2020	2846
Total	3478	4174	3486	4340	4366	4958
Under-sampling	824	1185	1162	1146	1455	1400

included in the hashtag (e.g., "#Baltimore" to "<hashtag> Baltimore"). Moreover, punctuation, symbols, and numbering were normalized.

3.5. Experimental design

To evaluate the potentials of the *MoralStrength* lexicon, we postulate the problem as a classification task. In particular, we employ the three approaches previously described, namely *Moral Freq*, *Moral Stats*, and *SIMON*, to predict the moral rhetoric in each of the aforementioned datasets (see Section 3.4).

In our experimental design, we include a basic Bag-of-Words (unigram) model providing a standardized way of obtaining a baseline in the computational linguistics field. We also report as a baseline the frequency counts employing the original MFD. We built a series of logistic regression models; firstly, we assess the predictive power of the unigrams, *Moral Freq*, *Moral Stats*, and *SIMON* lexicon methods alone. Then, we train logistic regression models concatenating the features extracted by the above approaches.[11] In this way, we examine the effective performance of

---

[9] The original corpus is available at https://github.com/t-davidson/hate-speech-and-offensive-language/tree/master/data.

[10] https://github.com/gsi-upm/gsitk.

[11] For replicability purposes, we have liberated the *MoralStrength* lexicon along with the implementation of the presented methods in a GitHub repository.[12]

[12] https://github.com/oaraque/moral-foundations.

both engineered and word embedding features in analyzing user-generated text. We also combine the unigrams to the proposed lexicon approaches described above. Hence, for each dataset and moral dimension, we train a series of logistic regression models following a 10-fold cross-validation scheme. We then report the F1-score as the evaluation metric per moral dimension since this is the once employed in the majority of the related studies.

To directly compare our proposed framework with the current state-of-the-art approach of Lin et al. [43], we replicated their same configuration. Namely, we perform over-sampling on the original dataset to overcome the highly imbalanced nature of the benchmark data (see Section 3.4). After over-sampling on the *Hurricane Sandy* data, we resulted with an average number of training examples, $N = 6128$, instead of the original dataset size, $N = 3478$ (see Table 2).

Since over-sampling implies "artificial" data samples, we propose an alternative methodology; more specifically, we performed under-sampling, which also deals with the issue of unbalanced classes, however, in doing so, it randomly excludes data points of the most populated class. In this way, for the *Hurricane Sandy* we had $N = 824$ data points (see Table 2. By reporting the score for both methods, we ensure the results are not biased by the technique used to address the class imbalance.

For all experiments, we report the performance in terms of F1-score, which is the metric also employed by Lin et al. [43], as well as the average F1-score over all moral dimensions. Moreover, to compare the improvement of the simplest model, which for this study we consider being the *Moral Freq* model, we employ the Friedman statistical test [50], which yields a ranking of the proposed method ordered by their performance. To obtain further insights on the statistical significance of our obtained results for the baseline model the Bonferroni–Dunn [50] post-hoc statistical test is performed with $\alpha = 0.05$.

4. Results and discussion

4.1. Evaluation of moral valence

After collecting the moral valence ratings, we assessed the quality of the crowdsourced data with an intrinsic evaluation. However, since the only dictionary currently available for MFT has binary annotations (vice/virtue), a direct comparison with it is not informative enough.

Hence, we evaluate the quality by (i) calculating inter-annotator agreement for the moral valence ratings, (ii) calculating the correlation between valence scores and the normative lexicon of Warriner et al. [48], and (iii) comparing binarized moral valence ratings with the gold standard given by the MFD. The results for all these tests are reported in Table 3.

To assess inter-annotator agreement we calculated Gwet's agreement coefficient (AC2) [51]. We opted for this measure since other, more common measures (e.g., Cohen's Kappa) require the number of annotators per element to be constant, and this is not the case for our data. Moreover, Gwet's coefficient can be weighted, meaning that annotators expressing close ratings will positively influence the coefficient score, and negatively impacted by scores that are far apart, a sensible feature for our dataset. Results for all the traits are in the "Moderate" to "Good" range (0.4–0.8), except for fairness (which had "Poor" agreement, 0.17). While this is positive, it also indicates that the task is not trivial and that some words might be hard to rate. The lower agreement of fairness led us to inspect the agreements for all traits manually, and we discovered that some annotators were particularly inaccurate. It was thus decided to discard some annotators, despite their ability to complete the crowdsourced experiment without failing the control questions. In particular, for the Authority trait,

Table 3

Measures of the quality of the collected ratings. The first column is the inter-annotator agreement for each moral dimension via Gwet's gamma with quadratic weighting metric. The second column is the correlation of the aggregate valence ratings and the gold standard of [48]. The last column is the agreement of the aggregate ratings (binarized) and the original Moral Foundations Dictionary, using Cohen's Kappa.

Moral trait	Inter-annotator	Warr correlation	MFD agreement
Authority	0.42	0.84	0.78
Care	0.65	0.95	0.91
Fairness	0.34	0.88	0.84
Loyalty	0.59	0.91	0.84
Purity	0.56	0.79	0.92

the annotator with the worst agreement was removed, improving the original AC2 of 0.41 to 0.42. For loyalty, the answer of one annotator was lost due to programmatic error (the result for one word is outside the range specified by the Likert scale) and was removed from the dataset (no effect on the agreement). In the case of fairness, we intervened more drastically and removed five annotators, plus one non-valid answer. The five discarded annotators were chosen due to them having a weak agreement with other annotators, and to inconsistent ratings (i.e., they gave the same score to antonyms that have opposite traits in the MFD gold standard, such as "honest" and "dishonest"). The inter-annotator agreement for valence ranges between 0.61 and 0.72, thus falling in the "Good" category for the set of words of every moral trait. This indicates, in general, that annotating valence is easier and less controversial than rating moral traits.

We also compared the aggregated values of valence ratings (i.e., the mean of all valence annotations for a word) with the gold scores provided by [48]. In this case, we report the results of the Pearson correlation, which ranges from 0.79 to 0.95, indicating once again that the crowdsourced annotation is of good quality, and that differences between annotators are within the acceptable range.

Finally, to be able to compare with the only gold standard for moral foundations, i.e., the Moral Foundations Dictionary, we binarized the aggregated annotations and excluded those whose average is 5 (the center of the Likert scale, meaning that the word is neither positive nor negative[13]). This way, we could calculate Cohen's kappa coefficient by comparing to the vice/virtue ratings of the MFD for the subset of words that exists in both datasets. The lowest agreement is for Authority, but also, in this case, the 0.78 value suggests that the annotations are generally reliable and entirely in line with the original MFD. It is perhaps worth noting that the agreement of fairness is quite good (0.84), despite the lower inter-annotator agreement of the collected ratings. This might indicate that, while the aggregate ratings are reliable (i.e., they fall in the correct side of the morality spectrum), there is a relatively high individual variation regarding where the words of that dimension should be placed.

4.2. Evaluation of MoralStrength lexicon

In this section, we assess the predictive power of the various approaches exploiting *MoralStrength* to analyze the moral rhetoric on the benchmark datasets described above. We confront the performance of the models against a series of baseline models. Initially, for each dataset, we report the performance of the model employing as features frequency counts from the MFD; this model shows how well the MFD alone would perform. Then,

---

[13] While it would be sensible to consider neutral a range instead of a single value, e.g., excluding everything in the interval 4.5–5.5, we wanted to avoid removing more words from the comparison.

we report the performance of the predictive model employing unigrams, which provide an assessment of the difficulty of the task itself and the State-Of-The-Art (SOTA) performance for each dataset when available.

We show the performance of the logistic regression models of increasing complexity, starting from the *Moral Freq, Moral Stats*, and *SIMON*, followed by aggregations of the above lexicons. For all experiments, we report the Friedman statistical test [50], which yields a ranking of the proposed methods ordered by their performance. To obtain further insights on the statistical significance of our obtained results to the baseline model the Bonferroni–Dunn [50] post-hoc statistical test is performed. Note that in this study, for the statistical significance test, we employed the *Moral Freq* with MFD model as a baseline model, and not the unigram one, since it is the one that infers on the simplest generated lexicon.

For the case of Hurricane Sandy we can see that across all moral dimensions, the model inferring on the aggregated unigram and *SIMON* features emerges as the best performing approach; with a statistically significant improvement of the average F1-score — 87.6 over 62.4 reported by Lin et al. [43] (see Table 4). In this case alone, we employed over-sampling to directly compare to the previous SOTA approach on the same dataset [43]. Interestingly, the highest score is obtained for "purity", which was reported being the most challenging moral dimension in the work of Lin et al. [43]. Examining each moral dimension separately, we note that our results are also consistently higher than the unigram model. The models that stand out are (i) unigrams + *SIMON* for fairness, loyalty, and purity, (ii) unigrams + *SIMON* + *Moral Freq* for care, purity, and neutral text, while (iii) unigrams + *SIMON* + *Moral Freq* + *Moral Stats* is the best performing models for authority and purity.

Table 5 reports the results of the evaluation for Hurricane Sandy when under-sampling is applied. Following this sampling approach, the results vary for oversampling (see Table 4), while the average overall performance slightly decreases (85.0% against 87.6% F1-score). Still, it is arguably preferable to perform under-sampling in comparison to over-sampling since in this way we avoid overfitting to the most prevalent outcome. Noteworthy is the fact that the best performing models are consistent between the over and under-sampling approaches. We also note that the importance of the statistical features regarding the moral valence of lemmas, exploited in the *Moral Freq* and *Moral Stats* lexicon methods, is more pronounced for all moral traits, for the oversampling technique. More precisely, the model inferring from unigrams together with the *Moral Stats* model, has a better performance in fairness, loyalty, and purity, while for care, the best performing model is the *SIMON* combined to the *Moral Freq* and *Moral Stats* Lexicons. Observing the obtained results, we conclude that combining lexicon-driven representations which take into consideration the moral valence, together with pure textual information (for instance, the unigrams), allows for a more robust and semantically meaningful representation. Despite the differences in the proposed approaches, comparing our approach to the study presented by Garten et al. [39], who also predicted the moral foundations on the *Hurricane Sandy* dataset, we note that their best performing model achieved 49.6% F1-Score, which is remarkably lower than the 88.2% reported here.

Next we present the performance of the models on *Baltimore Protest, All Lives Matter, Black Lives Matter, Davidson* and *2016 Presidential Election* datasets in Tables 4, 5, 6, 7, 8, 9 and 10, respectively. According to the previous discussion, we employ the under-sampling technique in all cases, always reporting the SOTA, as described in Hoover et al. [17].[14] Carefully comparing

the experimental results, we first note that unigrams provide a reasonably good baseline for all datasets. As expected, this method is shown to be a generally stable approach, even if the training samples are few. Next, we note that when we introduce the notion of moral strength to the basic unigram approach, results are steadily improved. Table 11, provides clear evidence to this statement; there, the Friedman test indicates that the best model overall is unigrams + *Moral Freq*, followed by unigrams + *Moral Stats*. Thus, introducing knowledge about moral valence, we can better predict the moral rhetoric in text.

This result differs from the statement made by Lin et al. [43], where they show that adding the features from the original lexicon, MFD, does not improve the score. Hence, we could argue that introducing the notion of moral valence the quality of the proposed lexicon, *MoralStrength*, improves the performance in text analysis as compared to the MFD. To safely conclude to this latest argument, we present a direct comparison of all datasets in Table 12. Here, we compare the performance of the original MFD versus our *MoralStrength* for all the proposed approaches. The reported scores are averaged over all moral values per dataset. As observed, the Friedman test indicates that the *SIMON* model with the proposed lexicon outperforms the rest. Hence, it is safe to assume that the newly introduced resource offers an improvement over the previous lexicon.

Moving to the model comparison, it can be seen that combining unigrams and *SIMON* does not generally improve the results of the classification. Interestingly enough, such combination does primarily improve the metrics when done in the over-sampling case (Table 4). In light of this contrast, and considering that both the unigrams and SIMON approach generate a large number of features, we hypothesize that combining large feature vectors leads to overfitting. As expected, an increase of the training data quantity improves the results, leading the unigrams + SIMON model to obtain better results.

To conclude, we observe that the performance trends are maintained; a unigram model is a robust approach, and adding information from *MoralStrength* improves the prediction performance. Noteworthy is the fact that the expression of moral sentiment can vary substantially according to the context. Variability in the model performances may also depend on the topic of discourse; the datasets employed for the evaluation include political left (e.g., BLM), right (ALM), both ideological poles (e.g., the Presidential election). They also include topics unrelated to political discourses (e.g., Hurricane Sandy). Moreover, the variability of training samples available for each trait may explain the differences in the model performance.

We believe that exploratory analysis will be useful for the ever-increasing studies on moral foundations since it presents a variety of approaches on how the moral lexicon we propose can be employed for the prediction of moral narratives from a text.

5. Conclusions

There is an ever-increasing interest in moral values understanding since they reflect our perception, attitudes, and opinion formation on critical societal issues. Moral values are expressed in user-generated content [52], and primarily through text. With the burst of social media data, emerges a unique opportunity of observing such behaviors in scale and as they happen. Recent

---

[14] Regarding *Baltimore Protest*, Rezapour et al. [45] reported higher accuracy with respect to Mooijman et al. [12]. Nevertheless, we cannot compare against

them for two reasons. First, the dataset is not the same; Rezapour et al. selected a subset from the original larger dataset [12] where annotations were automatically inferred by an algorithm, while our evaluation dataset originates by Hoover et al. [17], where annotations were manually assigned. Secondly, their evaluation metric is reported in terms of accuracy, while we use F1-Score as the majority of the related works. Thus, a direct comparison cannot be made.

Table 4
F1-Score of the proposed methods using over-sampling over *Hurricane Sandy* [43]. C/H: Care/Harm, F/C: Fairness/Cheating, L/B: Loyalty/Betrayal, A/S: Authority/Subversion, P/D: Purity/Degradation, NM: Non-moral, Avg.: Average. '*' and '†' mark that the approach significantly outperforms the MFD baseline and the SOTA, respectively. The model with the lowest rank is the one that outperforms the rest.

Approach	C/H	F/C	L/B	A/S	P/D	NM	Avg.	Rank
Baseline: Frequency MFD	56.3	59.2	61.8	54.4	63.1	66.4	60.2	14.4
unigrams	74.0	76.9	76.5	80.7	94.1	77.2	79.9	11.9
SOTA: Lin et al [43]	82.3	70.7	50.3	69.3	37.4	64.2	62.4	12.9
Moral Freq	61.4	58.2	61.9	56.0	62.1	63.4	60.5	14.4
Moral Stats	62.8	57.2	58.8	52.7	64.1	63.3	59.8	15.1
SIMON	79.6	82.3	77.1	86.0	98.1	84.2	84.5	6.4*
SIMON + *Moral Freq*	79.2	82.5	77.2	83.8	98.2	83.9	84.1	6.8*
SIMON + *Moral Stats*	79.2	82.2	77.0	84.0	98.2	83.9	84.1	7.6
SIMON + *Moral Freq* + *Moral Stats*	79.6	82.5	77.1	84.0	98.2	83.8	84.2	6.6*
unigrams + *Moral Freq*	75.3	77.7	77.2	81.2	95.5	77.8	80.8	9.7
unigrams + *Moral Stats*	73.5	77.6	76.7	81.3	95.7	77.9	80.5	10.8
unigrams + *Moral Freq* + *Moral Stats*	74.0	78.2	77.1	81.7	95.9	77.9	80.8	9.4
unigrams + SIMON	84.6	**85.6**	**81.2**	90.0	**98.9**	85.5	**87.6**	**2.0***†
unigrams + SIMON + *Moral Freq*	**85.1**	85.2	80.8	89.5	**98.9**	85.6	87.5	2.6*†
unigrams + SIMON + *Moral Stats*	84.9	85.4	80.4	90.0	98.8	85.2	87.5	3.3*†
unigrams + SIMON + *Moral Freq* + *Moral Stats*	85.0	85.4	80.8	**90.2**	**98.9**	85.2	**87.6**	2.1*†

Table 5
F1-Score of the proposed methods using under-sampling over *Hurricane Sandy* [43]. C/H: Care/Harm, F/C: Fairness/Cheating, L/B: Loyalty/Betrayal, A/S: Authority/Subversion, P/D: Purity/Degradation, NM: Non-moral, Avg.: Average. '*' marks that the approach statistically outperforms the baseline. The model with the lowest rank is the one that outperforms the rest.

Approach	C/H	F/C	L/B	A/S	P/D	NM	Avg.	Rank
Baseline: Frequency MFD	56.1	54.6	61.3	50.4	59.4	66.2	58.0	15.1
unigrams	78.1	88.8	85.7	**90.1**	66.5	**92.1**	83.5	5.3*†
SOTA: Hoover et al [17]	55.0	58.0	44.0	44.0	56.0	–	51.4	13.4
Moral Freq	65.2	56.4	61.6	51.7	54.6	69.3	59.8	14.6
Moral Stats	74.1	60.8	73.4	63.3	60.4	79.6	68.6	12.6
SIMON	76.9	76.3	77.2	75.0	73.5	72.8	75.3	10.7
SIMON + *Moral Freq*	77.2	79.6	79.4	75.6	70.7	73.4	76.0	10.6
SIMON + *Moral Stats*	77.3	80.8	80.2	74.6	72.2	76.0	76.8	10.1
SIMON + *Moral Freq* + *Moral Stats*	77.5	80.8	80.7	75.1	72.4	77.1	77.3	9.1
unigrams + *Moral Freq*	81.2	88.4	**86.8**	89.9	67.4	91.9	84.3	4.4*†
unigrams + *Moral Stats*	80.2	87.3	85.5	89.0	67.8	90.4	83.4	6.4*
unigrams + *Moral Freq* + *Moral Stats*	80.0	87.3	85.6	89.1	68.2	90.2	83.4	6.2*†
unigrams + SIMON	**83.8**	**92.0**	85.2	89.1	**74.7**	85.0	**85.0**	3.3*†
unigrams + SIMON + *Moral Freq*	83.1	91.5	86.0	88.1	73.2	85.7	84.6	3.6*†
unigrams + SIMON + *Moral Stats*	82.5	90.6	84.8	86.0	72.8	88.0	84.1	5.4*†
unigrams + SIMON + *Moral Freq* + *Moral Stats*	82.7	90.4	85.2	85.6	72.6	88.1	84.1	5.3*†

Table 6
F1-Score of the proposed methods using under-sampling over *Baltimore Protest* [17]. C/H: Care/Harm, F/C: Fairness/Cheating, L/B: Loyalty/Betrayal, A/S: Authority/Subversion, P/D: Purity/Degradation, NM: Non-moral, Avg.: Average. '*' and '†' mark that the approach significantly outperforms the MFD baseline and the SOTA, respectively. The model with the lowest rank is the one that outperforms the rest.

Approach	C/H	F/C	L/B	A/S	P/D	NM	Avg.	Rank
Baseline: Frequency MFD	59.3	64.4	64.3	50.2	54.7	64.1	59.5	14.9
unigrams	87.3	84.0	85.6	83.7	**93.2**	82.8	86.1	4.3*†
SOTA: Hoover et al [17]	33	47	25	25	15	–	29	13.9
Moral Freq	57.6	67.9	64.5	54.0	60.7	63.6	61.4	14.1
Moral Stats	66.5	72.5	71.0	63.7	65.6	68.1	67.9	12.3
SIMON	79.3	73.6	74.5	87.1	75.5	81.9	78.7	9.9
SIMON + *Moral Freq*	80.0	66.4	77.7	**87.5**	86.3	82.0	80.0	9.4
SIMON + *Moral Stats*	81.5	71.7	77.6	86.7	84.9	82.1	80.7	9.3
SIMON + *Moral Freq* + *Moral Stats*	81.5	73.8	77.6	85.8	84.9	82.1	81.0	8.7
unigrams + *Moral Freq*	88.1	81.4	85.2	85.4	91.9	84.0	86.0	4.1*†
unigrams + *Moral Stats*	88.0	**85.0**	**85.8**	86.7	90.5	83.5	86.6	3.4*†
unigrams + *Moral Freq* + *Moral Stats*	**88.5**	**85.0**	85.5	86.7	90.5	83.8	**86.7**	**3.1***†
unigrams + SIMON	86.0	68.5	84.8	81.7	87.8	85.1	82.3	7.6*
unigrams + SIMON + *Moral Freq*	86.4	65.4	84.2	81.2	90.5	85.1	82.1	8.2*
unigrams + SIMON + *Moral Stats*	86.6	72.1	84.7	82.9	86.4	**85.7**	83.1	6.8*†
unigrams + SIMON + *Moral Freq* + *Moral Stats*	87.1	72.1	84.8	83.3	86.4	**85.7**	83.3	6.1*†

developments in the computational linguistics domain, allow us to analyze automatically such data obtaining useful insights.

Operationalizing morality via the Moral Foundations Theory (MFT) [13], we propose a linguistic resource, *MoralStrength*, that aims at improving the only currently available dictionary, i.e., the MFD. More specifically, we contribute with a moral lexicon containing (i) a large number of lemmas, (ii) less radical and more frequently used lemmas, hence, improving its usability, and (iii) finally, containing a metric of moral valence for each lemma. *MoralStrength* contains approximately five times more lemmas

Table 7
F1-Score of the proposed methods using under-sampling over *ALM* [17]. C/H: Care/Harm, F/C: Fairness/Cheating, L/B: Loyalty/Betrayal, A/S: Authority/Subversion, P/D: Purity/Degradation, NM: Non-moral, Avg.: Average. '*' and '†' mark that the approach significantly outperforms the MFD baseline and the SOTA, respectively. The model with the lowest rank is the one that outperforms the rest.

Approach	C/H	F/C	L/B	A/S	P/D	NM	Avg.	Rank
Baseline: Frequency MFD	65.9	77.7	63.8	82.8	57.5	65	68.8	14.9
unigrams	74.0	**91.9**	88.1	90.7	92.0	**79.7**	**86.1**	**3.6***†
SOTA: Hoover et al [17]	67.0	76.0	62.0	63.0	39.0	–	61.4	14.3
Moral Freq	64.6	76.7	67.6	85.6	58.3	57.4	68.4	14.9
Moral Stats	64.4	85.4	85.3	**93.8**	70.6	60.3	76.6	11.6
SIMON	75.0	83.6	76.2	80.3	86.8	73.6	79.3	11.4
SIMON + *Moral Freq*	72.4	86.1	78.1	78.6	87.9	71.7	79.2	12.4
SIMON + *Moral Stats*	72.7	86.7	78.0	82.3	89.6	68.8	79.7	11.7
SIMON + *Moral Freq* + *Moral Stats*	72.3	87.7	79.9	81.7	90.1	69.2	80.2	11.1
unigrams + SIMON + *Moral Stats*	74.9	90.1	84.9	85.2	**93.1**	74.4	83.8	6.1*
unigrams + *Moral Freq*	74.5	90.4	88.3	90.7	89.8	78.2	85.3	4.6*†
unigrams + *Moral Stats*	70.8	90.9	89.1	90.7	92.8	73.2	84.6	5.9*†
unigrams + *Moral Freq* + *Moral Stats*	72.2	91.3	**89.3**	90.7	92.8	73.6	85.0	4.9*†
unigrams + SIMON	**75.6**	88.5	86.1	87.2	90.6	73.3	83.6	6.5*†
unigrams + SIMON + *Moral Freq*	74.8	88.4	83.9	86.5	90.4	73.7	82.9	8.1†
unigrams + SIMON + *Moral Stats*	74.9	90.1	84.9	85.2	**93.1**	74.4	83.8	6.1*†
unigrams + SIMON + *Moral Freq* + *Moral Stats*	**75.6**	90.2	84.9	86.5	**93.1**	74.2	84.1	5.0*†

Table 8
F1-Score of the proposed methods using under-sampling over *BLM* [17]. C/H: Care/Harm, F/C: Fairness/Cheating, L/B: Loyalty/Betrayal, A/S: Authority/Subversion, P/D: Purity/Degradation, NM: Non-moral, Avg.: Average. '*' and '†' mark that the approach significantly outperforms the MFD baseline and the SOTA, respectively. The model with the lowest rank is the one that outperforms the rest.

Approach	C/H	F/C	L/B	A/S	P/D	NM	Avg.	Rank
Baseline: Frequency MFD	68.0	84.3	89	89.3	82.1	70.6	80.5	15.4
unigrams	85.8	93.1	90.5	93.3	93.3	**81.6**	**89.6**	**3.0***†
SOTA: Hoover et al [17]	74.0	87.0	83.0	25.0	57.0	–	65.2	14.0
Moral Freq	66.2	86.2	88.8	92.3	80.8	70.9	80.9	13.9
Moral Stats	77.3	**93.4**	**92.8**	**96.6**	90.5	51.6	83.7	7.6*
SIMON	81.7	85.2	88.7	86.9	84.8	75.8	83.9	12.5
SIMON + *Moral Freq*	81.0	88.9	89.6	89.0	85.4	71.7	84.3	12.2
SIMON + *Moral Stats*	80.8	89.9	90.6	89.8	84.4	71.8	84.6	10.6
SIMON + *Moral Freq* + *Moral Stats*	80.9	89.8	90.6	90.1	87.2	72.5	85.2	9.4
unigrams + SIMON + *Moral Stats*	83.2	90.1	90.5	90.9	87.2	75.1	86.2	7.9*
unigrams + *Moral Freq*	86.1	91.6	89.6	93.1	92.3	79.9	88.8	4.6*
unigrams + *Moral Stats*	79.3	92.0	89.8	92.2	93.7	73.3	86.7	7.4*†
unigrams + *Moral Freq* + *Moral Stats*	80.3	92.3	89.8	92.2	**94.3**	73.3	87.0	6.8*
unigrams + SIMON	**88.2**	91.1	90.9	89.6	84.8	79.4	87.3	6.1*†
unigrams + SIMON + *Moral Freq*	87.7	88.7	90.5	90.1	90.3	76.4	87.3	6.7*†
unigrams + SIMON + *Moral Stats*	83.2	90.1	90.5	90.9	87.2	75.1	86.2	7.9*
unigrams + SIMON + *Moral Freq* + *Moral Stats*	84.0	91.4	90.5	90.7	87.6	75.5	86.6	6.9*

Table 9
F1-Score of the proposed methods using under-sampling over *Davidson* [17]. C/H: Care/Harm, F/C: Fairness/Cheating, L/B: Loyalty/Betrayal, A/S: Authority/Subversion, P/D: Purity/Degradation, NM: Non-moral, Avg.: Average. '*' and '†' mark that the approach significantly outperforms the MFD baseline and the SOTA, respectively. baseline. The model with the lowest rank is the one that outperforms the rest.

Approach	C/H	F/C	L/B	A/S	P/D	NM	Avg.	Rank
Baseline: Frequency MFD	36.5	33.3	39.4	36.1	38.7	38.0	37.0	14.8
unigrams	84.6	**91.3**	86.6	**77.3**	92.3	**56.5**	**81.4**	**4.4***†
SOTA: Hoover et al [17]	7.0	5.0	2.0	2.0	5.0	–	4.2	13.9
Moral Freq	39.6	33.3	39.9	39.1	37.9	39.6	38.2	14.3
Moral Stats	42.4	39.0	43.0	43.1	37.9	42.7	41.3	13.4
SIMON	84.8	85.8	87.2	72.0	89.3	53.8	78.8	11.2
SIMON + *Moral Freq*	85.4	85.8	87.7	71.1	89.8	54.3	79.0	9.4
SIMON + *Moral Stats*	85.7	85.8	**88.4**	72.4	89.8	54.5	79.4	7.6*
SIMON + *Moral Freq* + *Moral Stats*	85.7	85.8	**88.4**	72.4	89.8	54.4	79.4	7.8
unigrams + *Moral Freq*	84.6	90.6	87.6	76.2	92.3	56.0	81.2	4.9*†
unigrams + *Moral Stats*	85.4	**91.3**	87.4	77.0	90.6	55.8	81.2	4.9*†
unigrams + *Moral Freq* + *Moral Stats*	85.4	**91.3**	87.4	77.0	90.6	55.7	81.2	5.1*†
unigrams + SIMON	83.7	86.5	88.0	74.8	91.5	54.2	79.8	8.1
unigrams + SIMON + *Moral Freq*	85.3	86.5	88.0	75.2	**92.8**	54.9	80.5	6.0*†
unigrams + SIMON + *Moral Stats*	86.3	87.2	87.6	75.4	91.9	55.2	80.6	5.2*†
unigrams + SIMON + *Moral Freq* + *Moral Stats*	**86.4**	87.2	87.6	75.5	91.9	55.0	80.6	5.1*†

than the MFD, while at the same time providing with a moral valence, i.e., a quantitative assessment to characterize the lemmas' relationship with each moral dimension.

To explore the potentials of the moral lexicon in predicting the moral narrative in an unseen text, we generated three representations employing a series of feature extraction

Table 10

F1-Score of the proposed methods using under-sampling over *Election* [17]. C/H: Care/Harm, F/C: Fairness/Cheating, L/B: Loyalty/Betrayal, A/S: Authority/Subversion, P/D: Purity/Degradation, NM: Non-moral, Avg.: Average.'*' and '†' mark that the approach significantly outperforms the MFD baseline and the SOTA, respectively. baseline. The model with the lowest rank is the one that outperforms the rest.

Approach	C/H	F/C	L/B	A/S	P/D	NM	Avg.	Rank
Baseline: Frequency MFD	69.1	76.8	63.6	80.4	59.7	63.3	68.8	13.9
unigrams	86.6	**96.2**	90.7	87.0	**93.1**	63.2	86.1	6.1*†
SOTA	64.0	79.0	41.0	41.0	49.0	–	54.8	13.4
Moral Freq	68.2	70.7	66.1	83.6	59.8	59.0	67.9	14.1
Moral Stats	79.9	80.1	78.1	92.6	74.0	57.7	77.1	10.9
SIMON	82.2	78.7	83.0	72.5	64.8	69.0	75.0	11.9
SIMON + *Moral Freq*	81.5	82.6	80.9	74.0	68.0	69.6	76.1	10.9
SIMON + *Moral Stats*	81.5	83.4	83.7	74.2	64.9	68.9	76.1	11.0
SIMON + *Moral Freq* + *Moral Stats*	81.9	83.5	84.4	77.9	66.2	69.2	77.2	9.6
unigrams + *Moral Freq*	88.4	95.8	92.3	91.8	91.1	67.3	87.8	4.0*
unigrams + *Moral Stats*	88.1	94.4	**92.6**	**93.2**	92.7	65.8	87.8	4.1*
unigrams + *Moral Freq* + *Moral Stats*	**88.5**	95.1	92.5	**93.2**	93.0	66.4	**88.1**	**3.1*†**
unigrams + SIMON	88.1	93.6	92.3	85.9	81.4	70.9	85.4	5.6*†
unigrams + SIMON + *Moral Freq*	88.2	91.6	91.6	86.4	80.7	**71.6**	85.0	6.3*†
unigrams + SIMON + *Moral Stats*	87.4	93.1	92.1	88.1	82.7	70.1	85.6	5.9*†
unigrams + SIMON + *Moral Freq* + *Moral Stats*	87.4	93.1	91.8	88.4	83.1	70.2	85.7	5.4*†

Table 11

F1-Score of the proposed methods using under-sampling over all datasets. HS: Hurricane Sandy, BP: Baltimore Protest, ALM: All Lives Matter, BLM: Black Lives Matter, D: Davidson, PE: Presidential Election. '*' and '†' mark that the approach significantly outperforms the MFD baseline and the SOTA, respectively. The model with the lowest rank is the one that outperforms the rest.

Approach	HS	BP	ALM	BLM	D	PE	Rank
Baseline: Frequency MFD	58.0	86.1	80.5	80.5	37.0	68.8	13
unigrams	83.5	67.9	**89.6**	**89.6**	**81.4**	86.1	4.8*†
SOTA	51.4	86.6	65.2	65.2	4.2	54.8	13.8
Moral Freq	59.8	78.7	80.9	80.9	38.2	67.9	14.3
Moral Stats	68.6	80.0	83.7	83.7	41.3	77.1	12.7
SIMON	75.3	80.7	83.9	83.9	78.8	75	12.3
SIMON + *Moral Freq*	76.0	81.0	84.3	84.3	79.0	76.1	11.3
SIMON + *Moral Stats*	76.8	83.1	84.6	84.6	79.4	76.1	9.9
SIMON + *Moral Freq* + *Moral Stats*	77.3	86.0	85.2	85.2	79.4	77.2	8.4
unigrams + *Moral Freq*	84.3	86.6	88.8	88.8	81.2	87.8	**2.5*†**
unigrams + *Moral Stats*	83.4	**86.7**	86.7	86.7	81.2	87.8	4.3*†
unigrams + *Moral Freq* + *Moral Stats*	83.4	82.3	87	87	81.2	**88.1**	5.3*†
unigrams + SIMON	**85.0**	82.1	87.3	87.3	79.8	85.4	5.7†
unigrams + SIMON + *Moral Freq*	84.6	83.1	87.3	87.3	80.5	85	5.4†
unigrams + SIMON + *Moral Stats*	84.1	83.3	86.2	86.2	80.6	85.6	6.4
unigrams + SIMON + *Moral Freq* + *Moral Stats*	84.1	83.3	86.6	86.6	80.6	85.7	5.9†

Table 12

Average F1-Score of the proposed baselines using under-sampling over all datasets. '*' marks that the approach statistically outperforms the *Moral Freq. with the MFD lexicon* baseline. The model with the lowest rank is the one that outperforms the rest.

	Approach	HS	BP	ALM	BLM	D	PE	Rank
MFD	Moral Freq	58.0	59.5	68.8	80.5	37.0	68.8	5.5
	Moral Stats	64.8	67.3	74.1	80.2	39.2	75.6	4.0
	SIMON	74.0	**79.6**	79.1	82.9	78.4	74.0	2.3*
MoralStrength	Moral Freq	59.8	61.4	68.4	80.9	38.2	67.9	5.2
	Moral Stats	68.6	67.9	76.6	83.7	41.3	**77.1**	2.5*
	SIMON	**75.3**	78.7	**79.3**	**83.9**	**78.8**	75.0	**1.5***

techniques, including normalized lemmas frequencies, statistical features, and finally, semantic similarity based on word embeddings. We evaluated the machine learning framework on six benchmark datasets from the Twitter platform, the only available resources of linguistic data explicitly annotated for their moral content.

Interestingly, all our models improve the prediction performance with respect to the current state-of-the-art for all moral dimensions. The most prominent approaches – as indicated by the Friedman ranking – combine pure textual (e.g., unigrams) with lexicon-based representations (e.g., the *Moral Freq*, the *Moral Stats*, and the *SIMON*). Hence, we argue that moral lexicon can be successfully employed for moral values classification from a given text since when this information is considered, the models yield higher performance.

This study paves the way for further advancements in the moral text analysis, which is indeed an exciting field of study, both from the computational linguistics and the social sciences points of view. From a linguistic perspective, it would be interesting to explore how specific knowledge could be encoded in domain-oriented word vectors, allowing for the development of complex learning methods. Moreover, the word embedding representation based on moral similarity could be enhanced with the obtained assessments of moral valence, or even combined with sentiment features from the analyzed text. As for the social sciences, there are numerous issues where detecting the morals narrative can significantly improve our understanding of the peoples' dispositions in, for instance, controversial social phenomena as well as the evolution of opinions over time.

Acknowledgments

References

[1] E. Sagi, M. Dehghani, Moral rhetoric in Twitter: A case study of the US Federal Shutdown of 2013, in: Proceedings of the 35th Annual Meeting of the Cognitive Science Society (CogSci), vol. 36, 2014, pp. 1347–1352.

[2] C. Wolsko, H. Ariceaga, J. Seiden, Red, white, and blue enough to be green: Effects of moral framing on climate change attitudes and conservation behaviors, J. Exp. Soc. Psychol. 65 (2016) 7–19.

[3] M. Low, M.G.L. Wui, Moral foundations and attitudes towards the poor, Current Psychol. 35 (2016) 650–656.

[4] Y. Mejova, K. Kalimeri, Effect of values and technology use on exercise: Implications for personalized behavior change interventions, in: Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization, ACM, 2019, pp. 36–45.

[5] A.B. Amin, R.A. Bednarczyk, C.E. Ray, K.J. Melchiori, J. Graham, J.R. Huntsinger, S.B. Omer, Association of moral values with vaccine hesitancy, Nat. Hum. Behav. 1 (2017) 873–880.

[6] K. Kalimeri, M.G. Beiró, A. Urbinati, A. Bonanomi, A. Rosina, C. Cattuto, Human values and attitudes towards vaccination in social media, in: Companion Proceedings of the 2019 World Wide Web Conference, ACM, 2019, pp. 248–254.

[7] A. Miles, S. Vaisey, Morality and politics: Comparing alternate theories, Soc. Sci. Res. 53 (2015) 252–269.

[8] T. Grover, E. Bayraktaroglu, G. Mark, E.H.R. Rho, Moral and affective differences in us immigration policy debate on twitter, in: Computer Supported Cooperative Work, CSCW, 2019, pp. 1–39.

[9] M. Alizadeh, I. Weber, C. Cioffi-Revilla, S. Fortunato, M. Macy, Psychology and morality of political extremists: evidence from twitter language analysis of alt-right and antifa, EPJ Data Sci. 8 (1) (2019) 17.

[10] L.J. Skitka, G.S. Morgan, The social and political implications of moral conviction, Political Psychol. 35 (2014) 95–110.

[11] M.L. Vecina, D. Marzana, M. Paruzel-Czachura, Connections between moral psychology and intimate partner violence: Can ipv be read through moral psychology?, Aggress. Violent Behav. 22 (2015) 120–127.

[12] M. Mooijman, J. Hoover, Y. Lin, H. Ji, M. Dehghani, Moralization in social networks and the emergence of violence during protests, Nat. Hum. Behav. 2 (6) (2018) 389–396.

[13] J. Graham, J. Haidt, B.A. Nosek, Liberals and conservatives rely on different sets of moral foundations, J. Personal. Soc. Psychol. 96 (5) (2009) 1029.

[14] J. Haidt, J. Graham, When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize, Soc. Justice Res. 20 (1) (2007) 98–116.

[15] J. Haidt, C. Joseph, Intuitive ethics: How innately prepared intuitions generate culturally variable virtues, Daedalus 133 (4) (2004) 55–66.

[16] G.A. Miller, Wordnet: A lexical database for english, Commun. ACM 38 (11) (1995) 39–41.

[17] J. Hoover, G. Portillo-Wightman, L. Yeh, S. Havaldar, A.M. Davani, Y. Lin, B. Kennedy, M. Atari, Z. Kamel, M. Mendlen, et al., Moral foundations twitter corpus: A collection of 35k tweets annotated for moral sentiment, 2019.

[18] P. Stone, D. Dunphy, M. Smith, The General Inquirer: A Computer Approach To Content Analysis, MIT Press, 1966.

[19] P.D. Turney, Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews, in: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics, 2002, pp. 417–424.

[20] A. Muhammad, N. Wiratunga, R. Lothian, Contextual sentiment analysis for social media genres, Knowl.-Based Syst. 108 (2016) 92–101.

[21] E. Sulis, D.I.H. Farías, P. Rosso, V. Patti, G. Ruffo, Figurative messages and affect in twitter: Differences between# irony, # sarcasm and# not, Knowledge-Based Syst. 108 (2016) 132–143.

[22] C. Strapparava, R. Mihalcea, Learning to identify emotions in text, in: Proceedings of the 2008 ACM Symposium on Applied Computing (SAC), Fortaleza, Ceara, Brazil, 2008, pp. 1556–1560.

[23] B. Liu, Sentiment Analysis: Mining Opinions, Sentiments, and Emotions, Cambridge University Press, Cambridge, UK, 2015.

[24] S. Poria, A. Gelbukh, E. Cambria, A. Hussain, G.-B. Huang, Emosenticspace: A novel framework for affective common-sense reasoning, Knowl.-Based Syst. 69 (2014) 108–123.

[25] O. Araque, L. Gatti, J. Staiano, M. Guerini, Depechemood++: a bilingual emotion lexicon built through simple yet powerful techniques, IEEE Trans. Affect. Comput. (2019) 1, http://dx.doi.org/10.1109/TAFFC.2019.2934444.

[26] H.A. Schwartz, J.C. Eichstaedt, M.L. Kern, L. Dziurzynski, S.M. Ramones, M. Agrawal, A. Shah, M. Kosinski, D. Stillwell, M.E. Seligman, et al., Personality, gender, and age in the language of social media: the open-vocabulary approach, PLoS One 8 (9) (2013) e73791.

[27] T. Yarkoni, Personality in 100,000 words: A large-scale analysis of personality and word use among bloggers, J. Res. Pers. 44 (3) (2010) 363–373.

[28] R.L. Boyd, S.R. Wilson, J.W. Pennebaker, M. Kosinski, D.J. Stillwell, R. Mihalcea, Values in words: Using language to evaluate and understand personal values, in: Proceedings of the 9th International AAAI Conference on Web and Social Media (ICWSM), Oxford, UK, 2015, pp. 31–40.

[29] J. Chen, G. Hsieh, J.U. Mahmud, J. Nichols, Understanding individuals' personal values from social media word use, in: Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing, ACM, 2014, pp. 405–414.

[30] D. McAdams, J. Pals, A new big five: Fundamental principles for an integrative science of personality, Am. Psychol. 61 (3) (2006) 204–217.

[31] Y.R. Tausczik, J.W. Pennebaker, The psychological meaning of words: LIWC and computerized text analysis methods, J. Lang. Soc. Psychol. 29 (1) (2010) 24–54.

[32] S. Clifford, J. Jerit, How words do the work of politics: Moral foundations theory and the debate over stem cell research, J. Politics 75 (3) (2013) 659–671.

[33] L. Teernstra, P. van der Putten, L. Noordegraaf-Eelens, F. Verbeek, The morality machine: tracking moral values in tweets, in: Proceedings of the 15th International Symposium on Intelligent Data Analysis (IDA), Stockholm, Sweden, 2016, pp. 26–37.

[34] M. Dehghani, K. Sagae, S. Sachdeva, J. Gratch, Analyzing political rhetoric in conservative and liberal weblogs related to the construction of the "Ground Zero Mosque", J. Inf. Technol. Politics 11 (1) (2014) 1–14.

[35] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent Dirichlet allocation, J. Mach. Learn. Res. 3 (2003) 993–1022.

[36] E. Sagi, M. Dehghani, Measuring moral rhetoric in text, Soc. Sci. Comput. Rev. 32 (2014) 132–144.

[37] R. Kaur, K. Sasahara, Quantifying moral foundations from various topics on Twitter conversations, in: Proceedings of the IEEE International Conference on Big Data (BigData), Washington D.C. USA, 2016, pp. 2505–2512.

[38] O. Araque, G. Zhu, C.A. Iglesias, A semantic similarity-based perspective of affect lexicons for sentiment analysis, Knowl.-Based Syst. 165 (2018) 346–359.

[39] J. Garten, R. Boghrati, J. Hoover, K.M. Johnson, M. Dehghani, Morality between the lines: Detecting moral sentiment in text, in: Proceedings of the IJCAI 2016 Workshop on Computational Modeling of Attitudes (WCMA), New York, NY, USA, 2016.

[40] J. Garten, J. Hoover, K.M. Johnson, R. Boghrati, C. Iskiwitch, M. Dehghani, Dictionaries and distributions: Combining expert knowledge and large scale textual data content analysis, Behav. Res. Methods 50 (1) (2018) 344–361.

[41] J. Hoover, K. Johnson, R. Boghrati, J. Graham, M. Dehghani, Moral framing and charitable donation: Integrating exploratory social media analyses and confirmatory experimentation, Collabra: Psychol. 4 (1) (2018).

[42] J. Garten, B. Kennedy, J. Hoover, K. Sagae, M. Dehghani, Incorporating demographic embeddings into language understanding, Cogn. Sci. 43 (1) (2019).

[43] Y. Lin, J. Hoover, G. Portillo-Wightman, C. Park, M. Dehghani, H. Ji, Acquiring background knowledge to improve moral value prediction, in: Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), Barcelona, Spain, 2018, pp. 552–559.

[44] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780.

[45] R. Rezapour, S.H. Shah, J. Diesner, Enhancing the measurement of social effects by capturing morality, in: Proceedings of the Tenth Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, 2019, pp. 35–45.

[46] J. Graham, B.a. Nosek, J. Haidt, R. Iyer, S. Koleva, P.H. Ditto, Mapping the moral domain, J. Person. Soc. Psychol. 101 (2) (2011) 366–385.

[47] J.W. Pennebaker, The secret life of pronouns, New Scientist 211 (2828) (2011) 42–45.

[48] A.B. Warriner, V. Kuperman, M. Brysbaert, Norms of valence, arousal dominance for 13,915 English lemmas, Behav. Res. Methods 45 (4) (2013) 1191–1207.

[49] T. Davidson, D. Warmsley, M. Macy, I. Weber, Automated hate speech detection and the problem of offensive language, in: Eleventh International AAAI Conference on Web and Social Media, 2017.

[50] J. Demšar, Statistical comparisons of classifiers over multiple data sets, J. Mach. Learn. Res. 7 (Jan) (2006) 1–30.

[51] K.L. Gwet, Handbook of Inter-rater Reliability: the Definitive Guide to Measuring the Extent of Agreement Among Raters, Advanced Analytics, LLC, Gaithersburg, MD, USA, 2014.

[52] K. Kalimeri, M.G. Beiró, M. Delfino, R. Raleigh, C. Cattuto, Predicting demographics, moral foundations, and human values from digital behaviours, Comput. Hum. Behav. 92 (2019) 428–445.